

AGE- AND TIME-VARYING PROPORTIONAL HAZARDS MODELS FOR EMPLOYMENT DISCRIMINATION

BY GEORGE WOODWORTH AND JOSEPH KADANE

University of Iowa and Carnegie Mellon University

We use a discrete-time proportional hazards model of time to involuntary employment termination. This model enables us to examine both the continuous effect of the age of an employee and whether that effect has varied over time, generalizing earlier work [Kadane and Woodworth *J. Bus. Econom. Statist.* **22** (2004) 182–193]. We model the log hazard surface (over age and time) as a thin-plate spline, a Bayesian smoothness-prior implementation of penalized likelihood methods of surface-fitting [Wahba (1990) *Spline Models for Observational Data*. SIAM]. The nonlinear component of the surface has only two parameters, smoothness and anisotropy. The first, a scale parameter, governs the overall smoothness of the surface, and the second, anisotropy, controls the relative smoothness over time and over age. For any fixed value of the anisotropy parameter, the prior is equivalent to a Gaussian process with linear drift over the time–age plane with easily computed eigenvectors and eigenvalues that depend only on the configuration of data in the time–age plane and the anisotropy parameter. This model has application to legal cases in which a company is charged with disproportionately disadvantaging older workers when deciding whom to terminate. We illustrate the application of the modeling approach using data from an actual discrimination case.

1. Introduction. Federal law prohibits discrimination in employment decisions on the basis of age. There are two different bases on which a case may be brought alleging age discrimination. First, in a disparate impact case, the intent of the defendant is not at issue, but only the effect of the defendant’s actions on the protected class, namely, those forty or older. For example, a rule requiring new hires to have attained bachelor’s degrees after 1995 would be facially neutral, but would have the effect of preventing the hiring of older applicants. For such a case, data analysis is essential to see whether the data support disproportionate disadvantage to persons over

Received April 2009; revised January 2010.

Key words and phrases. Age discrimination, thin plate spline, smoothness prior, discrete proportional hazards, semiparametric Bayesian logistic regression.

This is an electronic reprint of the original article published by the Institute of Mathematical Statistics in *The Annals of Applied Statistics*, 2010, Vol. 4, No. 3, 1139–1157. This reprint differs from the original in pagination and typographic detail.

40 years of age with respect to whatever employment practices might be in question. Those practices might include hiring, salary, promotion and/or involuntary termination. A disparate treatment case, by contrast, claims intentional discrimination on the basis of age. Malevolent action, as well as intention, must be shown in a disparate treatment case. While statistics can address the defendant’s actions in a disparate treatment case, usually intent is beyond what data alone can address.

This paper uses a proportional hazards model as the likelihood [Cox (1972)]. Finkelstein and Levin (1994) used such a model using as dependent variable the positive part of $(age - 40)$ as an explanatory variable. Kadane and Woodworth (2004) treat age as a continuous variable, but do not model the response as a function of calendar time. This paper models both age and time continuously. This choice enables us to examine both the effect of age of an employee on employment decisions (our example uses involuntary terminations) and whether that effect has varied over time. Hence, there are two continuous variables, time and the age of the employee. In this way, the work here generalizes our earlier work [Kadane and Woodworth (2004)] that allowed continuous time, but reduced age to a binary variable (over 40/under 40). The analysis presented here allows us to address the extent to which a pattern or practice of age-based discrimination extends over a period of time. Proportional hazards regression is particularly suited to a pattern or practice case because it concerns the probability or odds of a person of a given age being involuntarily terminated relative to that of a person of another age (or range of ages), and hence directly addresses whether an older person is disproportionately disadvantaged.

We choose to use Bayesian inference because we find that it directly gives the probability that a person of a given age at a particular time is more likely to be fired than another person of a given other age at the same time. This contrasts with sampling-theory methods that give probabilities in the sample space, even after the sample is observed [Kadane (1990a)]. When combined with sensitivity analysis, Bayesian analysis permits us to assess the relative influence of the data and the model. We undertook the line of research in Kadane and Woodworth (2004) and in this paper to deal with temporally-sparse employment actions taken over a long time period. We particularly wanted to avoid the need to aggregate data into arbitrary time periods—months, quarters, years, etc.—in order to apply Cochran–Armitage type tests and the like.

2. Proportional hazards regression. The data required to analyze age discrimination in involuntary terminations comprise the beginning and ending dates of each employee’s period(s) of employment, that employee’s birth date, and the reason advanced by the employer for separation from employment (if it occurred). Table 1 is a fragment of the data analyzed in Section 3

TABLE 1
Flow data for the period June 1, 1989 to December 31, 1993

Birth date	Entry date	Separation date	Reason
⋮	⋮	⋮	⋮
3/1/1925	3/1/1961	6/1/1990	Vol ^a
4/9/1938	4/8/1961	8/17/1992	Vol
10/17/1934	4/5/1962	6/3/1992	Invol
12/9/1939	4/7/1962	12/18/1991	Invol
11/29/1932	5/29/1962	8/26/1989	Invol
9/5/1928	10/27/1962	6/12/1991	Vol
5/31/1941	1/12/1963	n/a	n/a
⋮	⋮	⋮	⋮

^a “Voluntary” termination includes death and retirement.

below. Data were obtained for all persons employed by a firm at any time between 06/07/1989 and 11/21/1993. The tenure of the last employee shown is right censored; that is, that employee was still in the work force as of 12/31/1993, and we are consequently unable to determine the time or cause of his or her eventual separation from the firm (involuntary termination, death, retirement, etc.).

2.1. *Overview.* The purpose of our statistical analysis is to determine how an employee’s risk of termination depends on his or her age and how the risk for employees of a given age changes with time. The idea is to estimate a surface such as the one in Figure 1 in such a way that it balances a penalty for infidelity to the data and for a penalty for a surface that is unrealistically “rough” [Gersch (1982)]. The result is a surface that is generally within the margins of sampling error but is also smooth. Smoothness, generally speaking, amounts to not having areas of high curvature (i.e., spikes, cliffs, buttes, sharp creases, etc.). The idea is to get a good fit to the data without sacrificing smoothness.

The mesh surface in Figure 1 is derived from a thin-plate spline model of the log odds (logit) of the probability of involuntary termination at a given time and age. The vertical axis shows the posterior median log-odds ratio of termination for employees of a given age on a given date relative to the weighted average rate for employees aged 39 years or younger on the same date (the legally unprotected class often used by statistical experts as a reference class for claims of disparate impact¹). The gray plane corresponds

¹Note, however, that Mr. Justice Scalia’s majority opinion in *O’Connor v. Consolidated Coin Caterers Corp.*, 517 U.S. 308 (1996) states that “though the prohibition is limited

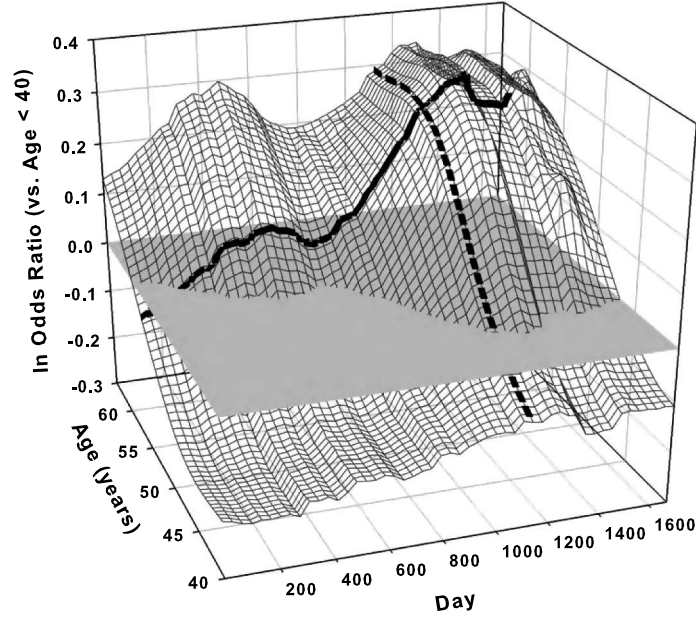


FIG. 1. *Smooth-model-derived log odds of termination relative to under-40 employees.*

to odds ratios equal to 1.00, indicating no age discrimination relative to the reference class; points above this plane exhibit discrimination. Although the underlying thin plate spline is smooth, the log-odds ratio surface is locally slightly rough because the observed numbers of employees in each age bin at the time of each termination were used as weights in computing the termination rate in the reference class.

The black ribbon in Figure 1 is the trajectory of the log-odds ratio over time for employees aged 56–57, and the dashed ribbon is the log-odds ratio as a function of age on day 1121 (05/30/92), the date of the involuntary termination of 57-year old plaintiff W1 in Case W described in Kadane and Woodworth (2004). The height of the surface at their intersection (0.297) is the posterior median log odds on the involuntary termination of 56–57 year-old employees relative to those under 40 on that date.

Figure 2 shows the posterior probability of age discrimination relative to under-40 employees as a function of age and date. Points above the gray plane represent dates and ages at which there was at least 70% posterior

to individuals who are at least 40 years of age, §631(a). This language does not ban discrimination against employees because they are aged 40 or older; it bans discrimination against employees because of their age, but limits the protected class to those who are 40 or older. The fact that one person in the protected class has lost out to another person in the protected class is thus irrelevant, so long as he has lost out *because of his age*.”

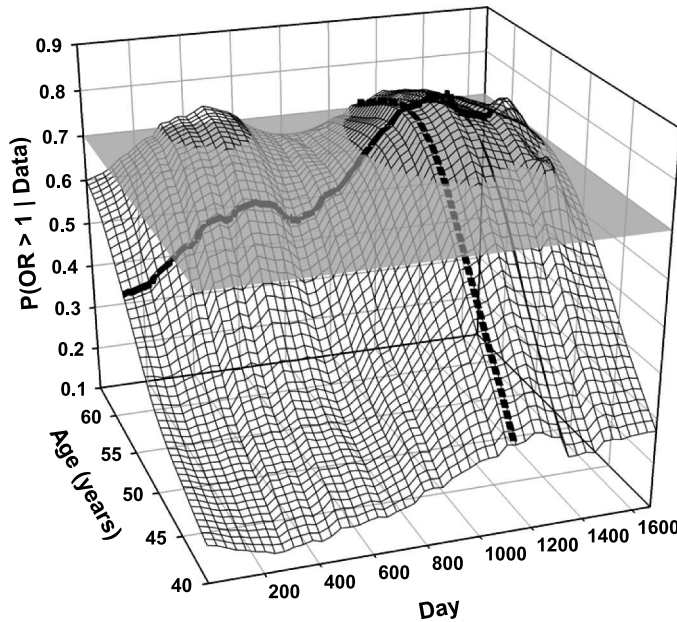


FIG. 2. Probability of age discrimination relative to under-40 employees.

probability of age discrimination. By itself, this would be comparatively weak evidence; however, Kadane (1990b), commenting on empirical research by Mosteller and Youtz (1990), suggests that this level of probability could, in standard usage, be said to make it “likely” that discrimination had occurred. The height of the surface at the intersection of the dashed and black ribbons (0.79) is the posterior probability that employees aged 56–57 were terminated at a higher rate compared to under-40 employees.

2.2. Proportional hazards models for time to event data. We are analyzing a group of individuals at risk for a particular type of failure (involuntary termination) for all or part of an observation period. The j th person enters the risk set at time h_j (either his/her date of hire or the beginning of the observation period) and leaves the risk set at time T_j either by failure (involuntary termination), or for other reasons (death, voluntary resignation, reassignment, retirement), or was still employed at the end of the observation period. The survival function $S_j(t) = P(T_j > t)$ is the probability that the j th employee is still employed at time t .

In practice, we rescale time and age to the unit interval $[0, 1]$ and, to make computations tractable, discretize each to a finite grid; $0 = t_0 < t_1 < \dots < t_p = 1$, $0 = a_0 < a_1 < \dots < a_r = 1$. Let p_{iw} be the conditional probability that employee (worker) w is terminated in the interval $(t_{i-1}, t_i]$ given the parameters and given that s(he) was in the workforce at time t_{j-1} . The discretized

data for this employee are $f_{iw}, \dots, f_{pw}; r_{iw}, \dots, r_{pw}$, where $r_{iw} = 1(0)$ if the employee was (not) in the work force (risk set) at time t_{i-1} , and $f_{iw} = 1(0)$ if the worker was (not) involuntarily terminated (fired) in that interval. The joint likelihood for all employees is $\prod_{w=1}^W \prod_{i=1}^p p_{iw}^{f_{iw}} (1 - p_{iw})^{r_{iw} - f_{iw}}$, where W is the total number of employees. Letting $a_w(t)$ denote the age of employee w at time t , we use the natural parametrization $\text{logit}(p_{iw}) = \beta(t_i, a_w(t_i))$, where $\beta(t, a)$ is a smooth function of time and age.

The aggregated data n_{ij} and x_{ij} are, respectively, the number of employees with ages in the interval $[a_{j-1}, a_j]$ at time t_i and the number of those who were terminated in that interval. At this level of aggregation, the likelihood is

$$(2.1) \quad l(\beta) = \prod_{i=1}^p \prod_{j=1}^r \exp(\beta_{ij} x_{ij} - n_{ij} \ln(1 + \exp(\beta_{ij}))),$$

where $\beta_{ij} = \beta(t_i, a_j)$. We assume that the grid is fine enough and the function smooth enough that variation of β within a grid cell is negligible. Changing the grid requires recomputing the cell counts, (n_{ij}, x_{ij}) and basis vectors defined below, which is fairly time consuming. We did a few runs with a grid roughly twice as fine (which quadrupled the run time and storage requirements) without observing substantive changes in the results; however, we focused our sensitivity analysis on varying the prior distribution of the smoothness parameter, which appeared to have much greater impact on the results. We compute the log-odds ratio at time t_i for employees aged a_j relative to unprotected employees (i.e., employees under age 40) as

$$(2.2) \quad \beta_{ij} - \text{logit} \left(\frac{\sum_{age_u \leq 40} n_{iu} p_{iu}}{\sum_{age_u \leq 40} n_{iu}} \right),$$

where age_u is age in years corresponding to scaled value a_u , and $\text{logit}(p_{ij}) = \beta_{ij}$.

2.3. Thin-plate spline smoothness priors. Likelihood measures fidelity to data (the larger the better); however, it does not incorporate our belief that the hazard ratio varies comparatively smoothly with time and age; this is provided by a roughness penalty (the smaller the better) that is subtracted from the log-likelihood

$$(2.3) \quad \frac{\lambda}{2} \iint \left[\left(\frac{\partial^2 \beta(t, a)}{\partial^2 t} \right)^2 + 2 \left(\frac{\partial^2 \beta(t, a)}{\partial t \partial a} \right)^2 + \left(\frac{\partial^2 \beta(t, a)}{\partial^2 a} \right)^2 \right] dt da.$$

The smoothness parameter, λ , weights the importance of smoothness relative to fidelity to noisy data (larger values of the smoothness parameter produces smoother fitted surfaces). However, there is no reason to expect

the log odds to be *isotropic*—equally smooth in time and age—and for that reason we assume that there is a rescaling $T = t/\sqrt{1+\rho^2}$, and $A = \rho a/\sqrt{1+\rho^2}$, such that the function $b(T, A) = \beta(T\sqrt{1+\rho^2}, A\sqrt{1+\rho^2}/\rho)$ is equally smooth (isotropic) in A and T . That is, the roughness penalty is

$$(2.4) \quad \frac{\lambda}{2} \iint \left[\left(\frac{\partial^2 b(T, A)}{\partial^2 T} \right)^2 + 2 \left(\frac{\partial^2 b(T, A)}{\partial T \partial A} \right)^2 + \left(\frac{\partial^2 b(T, A)}{\partial^2 A} \right)^2 \right] dT dA,$$

which reduces to the anisotropic roughness penalty,

$$(2.5) \quad \frac{\tilde{\lambda}}{2} \iint \left[\left(\frac{\rho^2}{1+\rho^2} \frac{\partial^2 \beta(t, a)}{\partial^2 t} \right)^2 + 2 \left(\frac{\rho}{1+\rho^2} \frac{\partial^2 \beta(t, a)}{\partial t \partial a} \right)^2 + \left(\frac{1}{1+\rho^2} \frac{\partial^2 \beta(t, a)}{\partial^2 a} \right)^2 \right] dt da,$$

where ρ is called the anisotropy parameter and $\tilde{\lambda} = \lambda \rho^3 / (1 + \rho^2)$. When $\rho = 1$ the surface is isotropic, and as $\rho \rightarrow \infty$ (or $\rho \rightarrow 0$), there is relatively less constraint on roughness in the age (or time) dimension.

It is interesting to compare this model to the earlier one of Finkelstein and Levin (1994), which is a special case of ours. In their case, our function $\beta(\cdot, \cdot)$ takes the form

$$\beta(t_i, a_w(t_i)) = (a_w(t_i) - 40)^+.$$

Since that function has zero second partial derivatives (except at 40, where they do not exist), their function imposes smoothness in our sense. One could think of this computationally as setting $\lambda = 0$.

Since the likelihood depends on the smooth function $\beta(t, a)$ only through the values β_{ij} , the roughness penalty is minimized for fixed β_{ij} when $\beta(t, a)$ is the interpolating thin-plate spline with values $\beta(t_i, a_j) = \beta_{ij}$. We have from Wahba [(1990), page 31, equation (2.4.9)] that there exist coefficients c such that the isotropic thin plate spline $b(T, A)$ can be represented as

$$(2.6) \quad b(T, A) = \sum_{ij} c_{ij} H(T - T_i, A - A_j) + l(T, A),$$

where $l(T, A)$ is an arbitrary linear function, $H(\mathbf{v}) = |\mathbf{v}|^2 \ln(|\mathbf{v}|) / (8\pi)$, and the coefficients c_{ij} satisfy the conditions $\sum_{ij} c_{ij} = \sum_{ij} t_i c_{ij} = \sum_{ij} a_j c_{ij} = 0$. Then the isotropic roughness penalty, equation (2.4), reduces to $\lambda \mathbf{c}' \mathbf{K}_\rho \mathbf{c}$, where \mathbf{c} is the vector of coefficients and \mathbf{K}_ρ is the $pr \times pr$ symmetric matrix with elements of the form $k_{ij,uv} = H(T_i - T_u, A_j - A_v) = H\left(\frac{(t_i - t_u)}{\sqrt{1+\rho^2}}, \frac{\rho(a_j - a_v)}{\sqrt{1+\rho^2}}\right)$.

To accommodate the constraints on vector \mathbf{c} , let \mathbf{P} be the projection onto the linear space orthogonal to the constraints so that $\mathbf{c} = \mathbf{P}\mathbf{c}$.

Finally, let $\mathbf{PK}_\rho\mathbf{P} = \mathbf{U}_\rho\mathbf{\Lambda}_\rho\mathbf{U}_\rho'$ be the spectral decomposition of $\mathbf{PK}_\rho\mathbf{P}$ and define the basis vectors \mathbf{B}_ρ as the nonzero columns of $\mathbf{U}_\rho\mathbf{\Lambda}_\rho^{1/2}$. It follows that the model for the vector of logits is

$$\begin{aligned} \beta &= \mathbf{K}_\rho\mathbf{c} + \mathbf{L}\tilde{\phi} \\ (2.7) \quad &= \mathbf{K}_\rho\mathbf{P}\mathbf{c} + \mathbf{L}\tilde{\phi} \\ &= \mathbf{PK}_\rho\mathbf{P}\mathbf{c} + (\mathbf{I} - \mathbf{P})\mathbf{K}_\rho\mathbf{P}\mathbf{c} + \mathbf{L}\tilde{\phi}, \end{aligned}$$

where β is the matrix with ij th row β_{ij} and the ij th row of matrix \mathbf{L} is $(1, t_i, a_j)$. But $\mathbf{I} - \mathbf{P}$ is the projection onto the column space of \mathbf{L} and, consequently, $(\mathbf{I} - \mathbf{P})\mathbf{K}_\rho\mathbf{P}\mathbf{c}$ can be absorbed into the linear term. Therefore, the model reduces to

$$\begin{aligned} \beta &= \mathbf{PK}_\rho\mathbf{P}\mathbf{c} + (\mathbf{I} - \mathbf{P})\mathbf{K}_\rho\mathbf{P}\mathbf{c} + \mathbf{L}\tilde{\phi} \\ (2.8) \quad &= \mathbf{U}_\rho\mathbf{\Lambda}_\rho^{1/2}(\mathbf{\Lambda}_\rho^{1/2}\mathbf{U}_\rho\mathbf{c}) + \mathbf{L}\phi \\ &= \mathbf{B}_\rho\delta + \mathbf{L}\phi, \end{aligned}$$

where $\delta = \mathbf{\Lambda}_\rho^{1/2}\mathbf{U}_\rho\mathbf{c}$ and $\mathbf{B}_\rho = \mathbf{U}_\rho\mathbf{\Lambda}_\rho^{1/2}$. Thus, for a given anisotropy, ρ , the columns of \mathbf{B}_ρ are basis vectors for the nonlinear part of the logit vector β .

The roughness penalty is $\lambda\mathbf{c}'\mathbf{K}_\rho\mathbf{c} = \lambda\mathbf{c}'\mathbf{PK}_\rho\mathbf{P}\mathbf{c} = \lambda\mathbf{c}'\mathbf{U}_\rho\mathbf{\Lambda}_\rho\mathbf{U}_\rho'\mathbf{c} = \lambda\delta'\delta$. The standard Bayesian interpretation of penalized likelihood estimation is that the penalty function is the log of the prior density of δ . Consequently, the components of that vector are a-priori independent and identically distributed normal random variables with precision λ . It follows that the prior conditional variance of β given λ , ρ and ϕ is

$$\begin{aligned} \text{Var}(\mathbf{B}_\rho\delta) &= \lambda^{-1}\mathbf{B}_\rho\mathbf{B}_\rho' \\ &= \lambda^{-1}\mathbf{PK}_\rho\mathbf{P} \end{aligned}$$

and, consequently, if \mathbf{d} is a vector such that $\mathbf{d}'\mathbf{L} = \mathbf{0}$, then

$$(2.9) \quad \text{Var}(\mathbf{d}'\beta) = \lambda^{-1}\mathbf{d}'\mathbf{K}_\rho\mathbf{d}.$$

The posterior distributions of λ and ρ are not well identified by the data and it is necessary to be somewhat careful about specifying their priors. However, the regression coefficients, ϕ , of the linear component do not influence smoothness, are well identified by the data, and can be given diffuse, normal prior distributions.

Viewing both time and age as continuous variables allows a more precise and general view of a firm's policy. However, due to the comparative sparseness of the data, some constraint on or penalty for roughness is needed to avoid an unrealistically rough model, unlike that depicted in Figure 1. It is, of course, possible to introduce discrete discontinuities into an otherwise

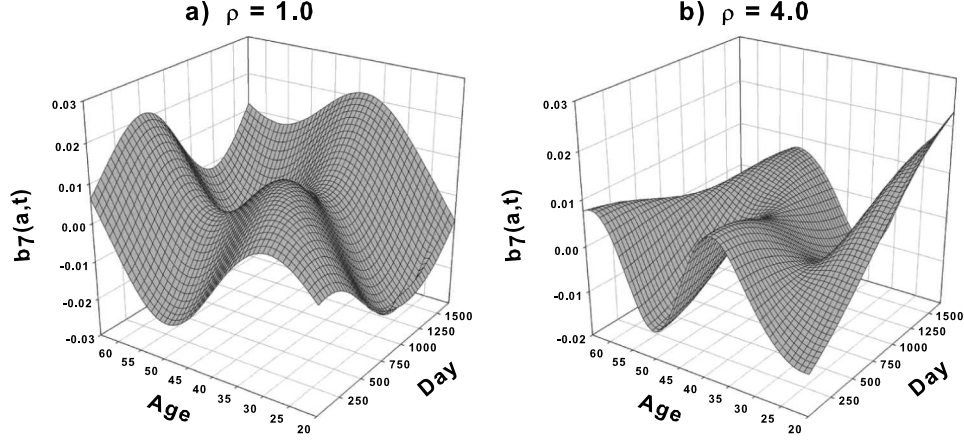
smooth model at time points where there is other evidence of a shift in employment practices [see, e.g., Figure 6 in Kadane and Woodworth (2004)]. However, we do not think that it is appropriate to “mine” for unknown numbers of discontinuities at unknown time points in the sparse data common in age-discrimination cases. Hence, it is necessary to smooth the data. The key parameters in doing so are smoothness and anisotropy. The smoothness parameter controls the average smoothness of the surface and the anisotropy parameter controls the relative degree of smoothing in the age and time coordinates.

3. Case W revisited. Over an observation period of about 1600 days the workforce at a firm was reduced by about two thirds; 103 employees were involuntarily terminated in the process. A new CEO took control at day 862, near the middle of the observation period. The plaintiff asserted that employees aged 50 and above were targeted for termination under the influence of the new CEO. Here we present a fully Bayesian analysis with smoothly time- and age-varying odds ratio. The personnel data were aggregated by status (involuntarily terminated, other) into one-week time intervals and two-year age intervals (20–21, 22–23, ..., 64–65). Figures 1 and 2 show posterior medians and posterior probabilities of age-related discrimination (i.e., of increased odds of termination relative to unprotected employees).

3.1. Forming an opinion about smoothness and anisotropy. The anisotropy parameter ρ governs the relative smoothness in time relative to age. This is clearly illustrated in Figure 3, which shows the seventh eigensurface (basis function) for (a) the isotropic case where there is about one cycle in either direction in contrast to (b) the anisotropic case $\rho = 4$ in which the surface is four times rougher in the age dimension (there are about 3 half cycles in the age dimension to about 3/4 of a half cycle in the time dimension).

In the context of employment discrimination, we think that, in terms of roughness of the logit, a 3-year age difference is about equivalent to a business quarter. Recalling that we have rescaled 1600 calendar days and a 45-year age span into unit intervals, a quarter is 0.056 and a three-year age interval is 0.067 of the unit interval, corresponding to anisotropy $\rho = 1.2$. We have found empirically that doubling or halving anisotropy has a fairly modest effect on surface shape; consequently, we used the prior distribution shown in Table 2, which has prior geometric mean 1.4.

As in our earlier analysis of this case [Kadane and Woodworth (2004)], we now derive a prior distribution for the smoothness parameter from our belief that the odds ratio on termination for a 10-year age difference are unlikely to change more than 15% over a business quarter. This implies that

FIG. 3. *Effect of anisotropy on the 7th basis function.*

a particular mixed difference is unlikely to exceed 0.15 in absolute value; that is, $\text{Prior}(|\Delta_t^2 \Delta_a \beta(t_0, a_0)| \leq 0.15)$ is large, where

$$\begin{aligned} \Delta_t^2 \Delta_a \beta(t_0, a_0) &= \beta(t_0 + 2d_t, a_0 + d_a) - 2\beta(t_0 + d_t, a_0 + d_a) + \beta(t_0, a_0 + d_a) \\ &\quad - \beta(t_0 + 2d_t, a_0) + 2\beta(t_0 + d_t, a_0) - \beta(t_0, a_0), \end{aligned}$$

where d_t is a rescaled half-quarter and d_a is a rescaled decade. We have from equation (2.9) that the prior distribution of $\Delta_t^2 \Delta_a \beta(t_0, a_0)$ is normal with mean zero and conditional variance, $\mathbf{d}'\mathbf{H}\mathbf{d}/\lambda = \mathbf{V}_\rho/\lambda$, where \mathbf{H} is the matrix with entries $H(T_i - T_{i'}, A_j - A_{j'})$, \mathbf{d} is the vector $(1, -2, 1, -1, 2, -1)$, $T_i = (t_0 + td_t)/\sqrt{1 + \rho^2}$, $i = 0, 1, 2$, and $A_j = \rho(a_0 + jd_a)/\sqrt{1 + \rho^2}$, $j = 0, 1$. Values of V_ρ are listed in Table 3.

The conditional prior distribution of the smoothness parameter given the anisotropy parameter is gamma with shape parameter and scale parameter selected so that $\text{Prior}(|\Delta_t^2 \Delta_a \beta(t_0, a_0)| \leq 0.15) = 1 - \alpha$ is large. To complete the derivation, we have, conditional on ρ , that

$$[\Delta_t^2 \Delta_a \beta(t_0, a_0)]^2 \sim V_\rho \cdot \frac{sc_\rho \Gamma(0.05)}{\Gamma(sh_\rho)} \sim V_\rho \cdot sc_\rho \frac{1 - \beta(sh_\rho, 0.05)}{\beta(sh_\rho, 0.05)},$$

TABLE 2
Prior distribution of the anisotropy parameter

ρ	8	4	2	1	0.5	0.25
Prior	0.08	0.16	0.26	0.26	0.16	0.08

Larger ρ -values favor smoothness in time.

TABLE 3
Prior variance $\times \lambda$ of $\Delta_t^2 \Delta_a \beta(t_0, a_0)$ and prior scale
parameter of λ

Anisotropy ρ	V_ρ	sc_ρ for $sh_\rho = 0.5$ and $\alpha = 0.05$
8	0.000383	5.04
4	0.000453	4.26
2	0.000492	3.93
1	0.000449	4.30
0.5	0.000332	5.81
0.25	0.000195	9.90

where, abusing the notation somewhat, we let $\Gamma(sh)$ denote an independent gamma-distributed random variable with shape parameter sh , and let $\beta(sh, 0.5)$ denote a beta-distributed random variable. Consequently, if

$$\text{Prior}([\Delta_t^2 \Delta_a \beta(t_0, a_0)]^2 \leq 0.15^2) = 1 - \alpha,$$

then

$$sc_\rho = \frac{0.15^2 \beta_\alpha(sh_\rho, 0.5)}{V_\rho(1 - \beta_\alpha(sh_\rho, 0.5))},$$

where $\beta_\alpha(sh_\rho, 0.5)$ is the α th quantile of the $\beta(sh_\rho, 0.5)$ distribution. The third column of Table 3 shows the values of the scale parameter, sc_ρ that we used to compute the surface in Figures 1 and 2.

3.2. Computing the posterior distribution. To estimate this model, we included enough basis vectors in the last row of equation (2.8) to account for at least 95% of the total roughness variance a priori (i.e., we included basis vectors accounting for 95% of the sum of the eigenvalues of K_ρ). We computed the posterior distribution of the probabilities of involuntary termination, and of the odds ratios relative to under-40 employees in each time-age bin using a program written in SAS IML language. For a given anisotropy value, ρ , we used the Metropolis-Hastings within the iteratively reweighted least squares algorithm proposed by Gamerman (1997) to separately update the logistic regression coefficient vectors ϕ and δ , and a Gibbs step to update the smoothness parameter, λ . Anisotropy values were chosen from the six shown in Table 2, where, beginning with an arbitrary initial value, we attempted a jump from the current anisotropy value to an adjacent value with transition probabilities from the 6×6 doubly stochastic matrix shown in Table 4. Letting current parameter values be δ , ϕ , λ , and ρ , we attempt a reversible jump, $\rho \rightarrow \tilde{\rho}$. We then propose values $\tilde{\phi} = \phi$, and $\tilde{\lambda} = \rho \cdot sc / \tilde{sc}$, where sc and \tilde{sc} are scale parameters from Table 3 corresponding to ρ and $\tilde{\rho}$, respectively. Finally, we generate a proposal for $\tilde{\delta}$ as follows.

Let $\beta = \mathbf{B}_\rho \delta + \mathbf{L} \cdot \phi$ be the current logit vector and let \mathbf{p} be the current vector of termination probabilities in time-age bins [i.e., $\text{logit}(\mathbf{p}) = \beta$] and let $\mathbf{q} = 1 - \mathbf{p}$. Let vectors \mathbf{n} and \mathbf{y} be the numbers at risk and terminated in the time-age bins. Then, $\tilde{\delta}$ is proposed from the multivariate normal distribution with precision $\tilde{\Pi} = [\tilde{\lambda} + \mathbf{B}'_{\tilde{\rho}} \mathbf{n} \mathbf{p} \mathbf{q} \mathbf{B}_{\tilde{\rho}}]$ and mean $\tilde{\mu} = \tilde{\Pi}^{-1} \mathbf{B}'_{\tilde{\rho}} \mathbf{n} \mathbf{p} \mathbf{q} \cdot \hat{\mathbf{y}}$, where \mathbf{B}_ρ is the matrix of basis vectors corresponding to anisotropy ρ , as defined in the paragraph after equation (2.8), and $\hat{\mathbf{y}} = \mathbf{B}_\rho \delta + (\mathbf{y} - \mathbf{p})/\mathbf{p} \mathbf{q}$. The proposal is accepted with probability

$$\begin{aligned} \alpha &= \min \left[1, \frac{p(\tilde{\rho}) p(\tilde{\lambda} | \tilde{\rho}) p(\delta | \tilde{\lambda}) l(\tilde{\beta})}{p(\rho) p(\lambda | \rho) p(\delta | \lambda) l(\beta)} \cdot \frac{p(\tilde{\rho} \rightarrow \rho) q(\delta | \tilde{\lambda}, \tilde{\delta}, \phi)}{p(\rho \rightarrow \tilde{\rho}) q(\tilde{\delta} | \lambda, \delta, \phi)} \cdot \frac{\partial \tilde{\lambda}}{\partial \lambda} \right] \\ &= \min \left[1, p(\tilde{\rho}) \tilde{\lambda}^{\tilde{q}/2} \exp \left(-\frac{1}{2} \tilde{\lambda} \delta' \tilde{\delta} \right) l(\tilde{\beta}) \right. \\ &\quad \times |\Pi|^{0.5} \exp \left(-\frac{1}{2} (\delta - \mu') \Pi (\delta - \mu)' \right) \\ &\quad \left. / \left(p(\rho) \lambda^{q/2} \exp \left(-\frac{1}{2} \lambda \delta' \delta \right) l(\beta) \right. \right. \\ &\quad \left. \left. \times |\tilde{\Pi}|^{0.5} \exp \left(-\frac{1}{2} (\tilde{\delta} - \tilde{\mu})' \tilde{\Pi} (\tilde{\delta} - \tilde{\mu}) \right) \right) \right], \end{aligned}$$

where $l(\beta)$ is the likelihood function [equation (2.1)], q and \tilde{q} are the ranks of B_ρ and $B_{\tilde{\rho}}$, and μ and Π are the mean and precision of the reverse proposal [Green (1995)].

3.3. Sensitivity analysis. It is a good statistical practice to investigate whether and to what extent the results of an analysis are sensitive to the prior distribution. That means in this case investigating the influence of the prior distribution of the smoothness and anisotropy parameters. Figures 1 and 2 above are based on our preferred prior distribution as specified in Tables 2 and 3. In Figure 3 we compare Figure 1 (a) with an analysis (b) in which the scale parameters in Table 4 are multiplied by 10, decreasing the roughness penalty by a factor of 10 and producing a substantially rougher surface. Figure 5 shows the effect of this variation on the probability of discrimination.

3.4. Identification of the anisotropy parameter. Table 5 shows the marginal posterior distribution of the anisotropy parameter for the preferred prior distribution of the smoothness parameter (Table 3). The posterior probability $P(\rho | \text{Data})$ is the observed rate of sampler visits to value ρ of the anisotropy parameter in 19,000 replications, the marginal likelihood

TABLE 4
Jump proposal probabilities for the anisotropy parameter

Anisotropy	8	4	2	1	0.5	0.25
8	0.9	0.1				
4	0.1	0.8	0.1			
2		0.1	0.8	0.1		
1			0.1	0.8	0.1	
0.5				0.1	0.8	0.1
0.25					0.1	0.9

is $P(\rho|Data)/P(\rho) \propto P(Data|\rho)$, and $p_{0.025}$ and $p_{0.975}$ are nominal Monte-Carlo error bounds computed on the assumption that the observed rate has a binomial distribution.

It is clear from the marginal likelihood that the data carry information about anisotropy and, in particular, that models with large values of ρ (i.e., which are very rough in the time dimension) are disconfirmed by the data. However, high levels of smoothness in the time dimension are not disconfirmed by data and apparently must be discouraged by the prior. Because of this, we investigated the effect of a prior that forces more smoothness in the time dimension.

In Figure 6 we altered the prior distribution for the anisotropy parameter to favor smoothness in the time dimension (Table 6). In this case the prior geometric mean of the anisotropy parameter is about 4, meaning that we think that, in terms of roughness of the log odds on termination, a decade of age is about equivalent to a business quarter (see Section 3.1). Evidence

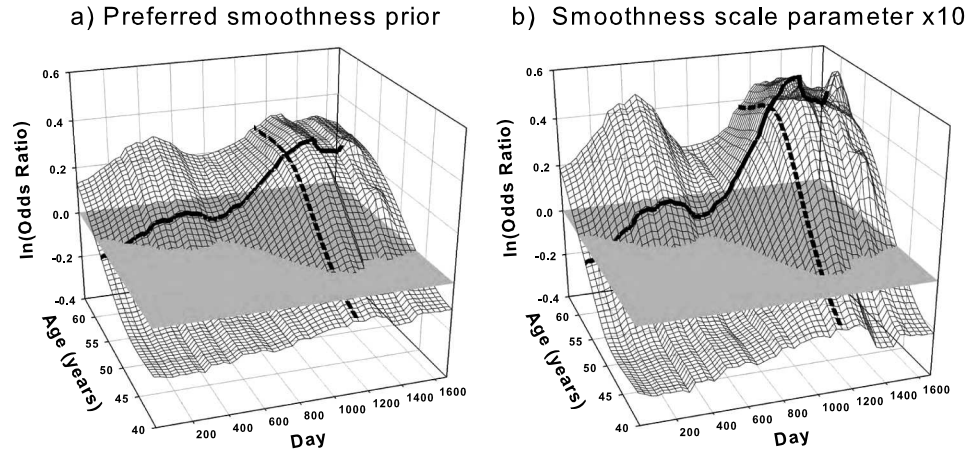


FIG. 4. *Effect of the smoothness prior on the log odds ratio.*

TABLE 5
Posterior distribution and marginal likelihood of the anisotropy parameter

ρ	Prior	Posterior ^a			Marginal likelihood		
		$P(\rho Data)$	$p_{0.025}$	$p_{0.975}$	$\propto P(Data \rho)$	$p_{0.025}$	$p_{0.975}$
8	0.08	0.122	0.12	0.13	1.53	1.47	1.61
4	0.16	0.231	0.22	0.24	1.44	1.40	1.50
2	0.26	0.286	0.28	0.30	1.10	1.07	1.14
1	0.26	0.217	0.21	0.23	0.83	0.81	0.87
0.5	0.16	0.101	0.10	0.11	0.63	0.61	0.67
0.25	0.08	0.043	0.04	0.05	0.54	0.50	0.59

^a $p_{0.025}$ and $p_{0.975}$ are Monte-Carlo error bounds (see text).

of discrimination in the plaintiff's case (the intersection of the dashed and black ribbons) is slightly stronger for the prior that forces more smoothness in the time dimension; $P(OR > 1|Data)$ is about 0.79 for the preferred prior (a) and about 0.83 for the more time-smoothing prior (b).

Although the analysis in panel (b) is more favorable to the plaintiff, we think it would be less persuasive to the trier(s) of fact (judge or jury) since it does not seem to distinguish between the periods before and after the arrival of the new CEO (day 862).

3.5. *Previous analyses of case W.* The plaintiff who was between 50 and 59 years of age was one of 12 employees involuntarily terminated on day 1092. He brought an age discrimination suit against the employer under the theory that the new CEO had a pattern of targeting employees aged 50 and above for termination.

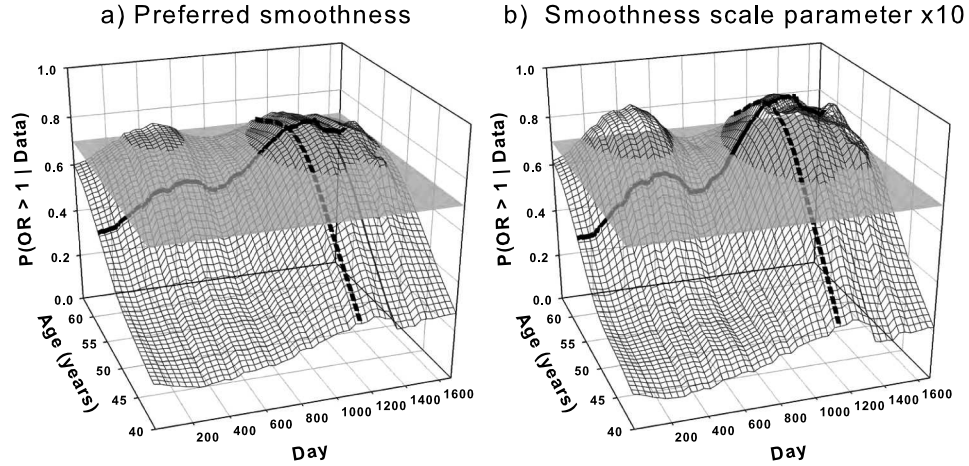


FIG. 5. *Effect of the smoothness prior on the posterior probability of discrimination.*

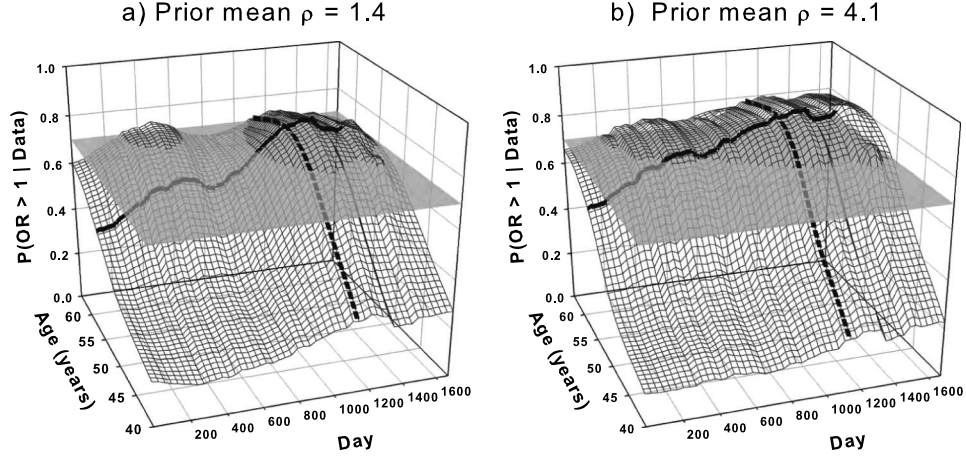


FIG. 6. *Effect of the anisotropy parameter on the posterior probability of discrimination.*

In the original case, the plaintiff’s statistical expert tabulated involuntary termination rates for each calendar quarter and each age decade. He reported that, “[Involuntary] separation rates for the [period beginning at day 481] averaged a little above three percent of the workforce per quarter for ages 20 through 49, but jumped to six and a half percent for ages 50 through 59. The 50–59 year age group differed significantly from the 20–39 year age group (signed-rank test, $p = 0.033$, one sided).” The plaintiff alleged and the defendant denied that the new CEO had vowed to weed out older employees. The case was settled before trial.

In a subsequent re-analysis [Kadane and Woodworth (2004)], we employed a proportional hazards model with separate, smoothly time-varying log hazard ratios for ages 40–49, and 50–64, with ages 20–39 as the reference category. Thus, the log hazard ratio was smooth over time but piecewise constant over age; Figure 7 is reproduced with permission from that paper. Our preferred model, represented by the solid curves, had prior mean smoothness 0.007. For this prior the posterior probability of age-discrimination in the case of Plaintiff W1 was 0.842.

The model depicted in Figure 7 has two explanatory variables for age, an indicator variable for age in the range 40–49 and an indicator variable

TABLE 6
Alternate prior distribution of the anisotropy parameter

ρ	8	4	2	1	0.5	0.25
Prior	0.5	0.25	0.125	0.0625	0.03125	0.03125

Larger ρ -values favor smoothness in time.

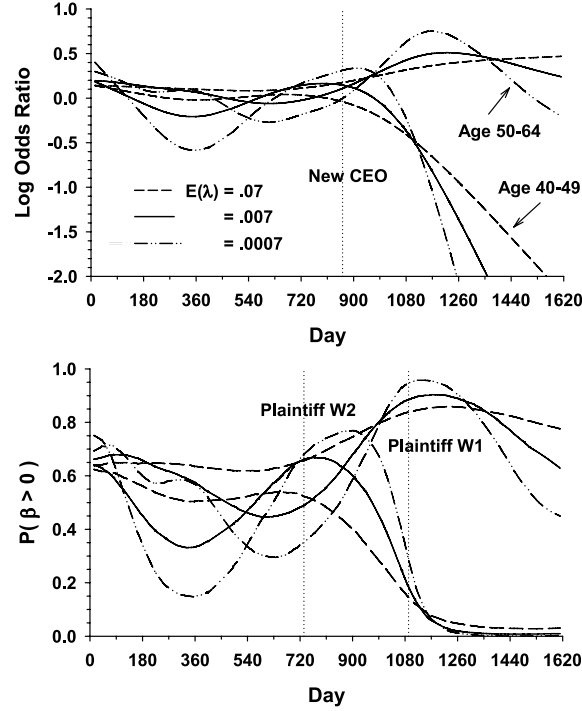


FIG. 7. Smooth by piecewise constant proportional hazards model.

for age 50 and above (there are no employees 65 and over in the data set). The likelihood model was proportional hazards regression with smoothly time-varying coefficients for the two explanatory variables. Three analyses are shown here with different prior means for the smoothness parameter, λ . The upper panel shows posterior means of the proportional-hazards regression coefficients as functions of time and the smoothness parameter. As suggested in the figure, the regression coefficients are interpretable as instantaneous log-odds ratios with unprotected, under-40, employees as the reference category. The second panel presents posterior probabilities that the two regression coefficients are positive; that is, that the termination rate is higher for the protected subclasses compared to the unprotected class. For example, at the time of plaintiff W2's termination, the posterior probability exceeds 80% that employees age 50 and above had a higher risk of termination than the protected class.

A second plaintiff, W2 aged 60 terminated on day 733, also brought an age-discrimination suit on the theory that employees aged 60 and above were disproportionately targeted at the time of his termination. On that day three of eight employees (37.5%) aged 60 and up were terminated compared to 15 of 136 (11.0%) employees terminated out of all other age groups (one-sided

Fisher exact test $p = 0.0530$). In our re-analysis the posterior probability of age discrimination against employees aged 50–64 was about 50% but did not distinguish between employees aged 50–59 and 60–64. Our second re-analysis reported in this paper remedies that deficiency and gives a more detailed picture of the impact of age on the risk of discrimination; in particular, for our preferred prior, the posterior probability of age discrimination against 60-year old employees on day 733 is about 65% but is only about 37% for 50-year old employees.

3.6. *Summary.* Table 7 summarizes the results of the three analyses of case W for each of the two plaintiffs. In the first, classical, analyses for Plaintiff W1, it is assumed that each employee in the age groups 20–39 and 50–59 has the same chance of being involuntarily terminated (i.e., fired) in each quarter-year after day 481. The test of significance calculates the probability of obtaining data as or more extreme than that observed were it true that persons in these two age groups have the same chance of being fired in any given quarter. The classical analysis for plaintiff W2 is somewhat different, in that it focuses solely on what happened on the day that W2 was fired. It conditions on both the age distribution of the workforce at the time (eight of 144 employees 60 years old or older) and the number fired (18) and computes the probability of three or more of the eight older employees being fired, if employees were equally likely to be fired.

TABLE 7
Summary of three analyses of Case W

Analysis	Method	Figure of merit	Treatment of age	Age \times time interaction	Plaintiff	
					W1	W2
Original expert's report	Frequentist	p -value	categorical: 40-up	none	0.033	0.053
Kadane and Woodworth (2004)	Bayesian	probability of disproportional disadvantage	categorical: 40–49, 50–64	smooth	0.84	0.50
				smooth/w discontinuity at day 862	0.88	0.49
This paper	Bayesian	probability of disproportional disadvantage	smooth	smooth	0.65	0.37
Anonymous referee of this paper	Cox regression	p -value, OR, and 90% LCL	linear above 40	none but restricted to day 1000 up	p : 0.041 OR: 2.04 LCL: 1.01	n/a

The second analysis is based on a model for the log odds of being fired that is continuous in time but still assumes constancy in age categories. The analysis of this paper relaxes this latter assumption, and allows smoothness in both age and time. In both Bayesian analyses, the probability computed is that an employee of a given age was more likely to be fired at a particular time than was an employee in the unprotected 20–39 age group.

Although the classical analyses are computing probabilities in the sample space while the Bayesian analyses are computing probabilities in the parameter space, the stronger effect here appears to be that as the assumptions get less rigid, there is less certainty that these plaintiffs’ cases were meritorious, as Table 7 shows. In view of the tendency of Bayesian analyses to draw estimates toward each other, this is perhaps not too surprising.

4. Discussion. In a nonhierarchical model, the effect of the prior can be isolated by separately reporting the likelihood function and the prior distribution. In particular, if the parameter space is divided into two disjoint subsets, the likelihood ratio and the prior odds suffice. However, in a hierarchical model such as this one, such a separation is not possible. For this reason, we have reported the results of changing our prior directly, in Sections 3.3, 3.4 and 3.5.

We have presented a global analysis of involuntary terminations that incorporates all of the data but reflects fine-grained variations over time and age of employee. The results are somewhat sensitive to assumptions about prior distribution of the smoothness parameter, although not enough to materially alter the strength of evidence supporting the plaintiff’s discrimination claim in Case W. This analysis, in our view, casts new light on the apparent patterns in coarser-grained descriptive presentations that might be easier for nonspecialists to grasp.

Our intent is to develop a methodology that does not require complex assumptions about the relationship between time, age and risk of termination. Indeed, the only structural assumption is smoothness and the only prior opinion required has to do with the degree of smoothness. We have suggested how that prior opinion could be elicited by considering how rapidly the risk of termination is likely to change over a business quarter and over a decade of age. A referee described our analysis as “staggeringly complex” and “shuddered to think what a judge or jury would make of this approach.” All statistical analyses are “staggeringly complex” to most laypersons. We think our responsibility as statisticians (and experts in court) is to present our best analysis of the data, and to explain it as best as we can.

A global analysis such as this one is more powerful and more appropriate than analyzing subsets of the data, perhaps in the form of individual termination waves or individual business quarters, and more appropriate than analyzing coarse aggregations such as employees aged 40 and above

compared to younger employees. The fallacy of subdividing the data is that such analyses implicitly assume that there is no continuity in the behavior of a firm and no difference in treatment of employees of different ages within the same broad age category (40 and older). We believe that the appropriate approach to possible inhomogeneities of the age effect is to incorporate them in a global model—see, for example, our discussion of Gastwirth’s (1992) analysis in *Valentino v. United States Postal Service* [Gastwirth (1992), Kadane and Woodworth (2004)].

Finally, it has not escaped our notice that our analysis of Case W has made it clear that only a subgroup of older employees, centered around the peak at day 1275 and age 54–55, has even moderately strong statistical evidence to support a claim of age discrimination. We believe that this is precisely the information that the court needs in order to determine how an award (if any) should be distributed among members of a certified class.

SUPPLEMENTARY MATERIAL

Supplement A: Employment — Case W (DOI: [10.1214/10-AOAS330SUPPA](https://doi.org/10.1214/10-AOAS330SUPPA); .txt). Data from two cases described in the paper “Hierarchical models for employment decisions,” by Kadane and Woodworth. A constant number of days has been subtracted from each date to preserve confidentiality.

Supplement B: Code for calculations (DOI: [10.1214/10-AOAS330SUPPB](https://doi.org/10.1214/10-AOAS330SUPPB); .zip).

REFERENCES

- COX, D. R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. [MR0341758](#)
- FINKELSTEIN, M. O. and LEVIN, B. (1994). Proportional hazards models for age discrimination cases. *Jurimetrics Journal* **34** 153–171.
- GAMERMAN, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* **7** 57–68.
- GASTWIRTH, J. (1992). Employment discrimination: A statistician’s look at analysis of disparate impact claims. *Law and Inequality: A Journal of Theory and Practice* **XI** Number 1.
- GERSCH, W. (1982). Smoothness priors. In *Encyclopedia of Statistical Sciences* (S. Kotz, N. L. Johnson and C. B. Read, eds.) **8** 518–526. Wiley, New York.
- GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. [MR1380810](#)
- KADANE, J. B. (1990a). A statistical analysis of adverse impact of employer decisions. *J. Amer. Statist. Assoc.* **85** 925–933.
- KADANE, J. B. (1990b). Comment: Codifying chance. *Statist. Sci.* **5** 18–20.
- KADANE, J. B. and WOODWORTH, G. G. (2004). Hierarchical models for employment decisions. *J. Bus. Econom. Statist.* **22** 182–193. [MR2049920](#)
- MOSTELLER, F. and YOUTZ, C. (1990). Quantifying probabilistic expressions. *Statist. Sci.* **5** 2–12. [MR1054855](#)

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.
[MR1045442](#)

DEPARTMENT OF STATISTICS
AND ACTUARIAL SCIENCES
UNIVERSITY OF IOWA
241 SCHAEFER HALL
IOWA CITY, IOWA 52240
USA
E-MAIL: george-woodworth@uiowa.edu

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
232 BAKER HALL
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: kadane@stat.cmu.edu